# Red Storm Update
# HPC User Forum

Erik P. DeBenedictis

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy under contract DE-AC04-94AL85000.

# Outline

- Project Organization
- Processor
- Network and Network Topology
- Light Weight Kernel (LWK)
- Reliability, Availability and Serviceability (RAS)

# Project Organization

- ASCI Red was very successful
- Red Storm RFQ very nearly ASCI Red sped up by Moore's Law (7x)
- Cray is selling Red Storm to Sandia as a custom product
  - However, Sandia is supplying key expertise for this specific architecture to Cray, and
  - Sandia supplying a major part of the systems software to Cray for integration into Cray's product
- This organization is working

# Processor

- **Sandia did not specify a processor, but concurs with Cray that the Opteron is a very good choice**
- **Sandia conducted an evaluation of many available processors**
    - **Considered overall ability of a processor to integrate into a system**
    - **Specifically considered FLOPS, memory bandwidth, I/O bandwidth, power consumption**
    - **Ran benchmarks of top Sandia/ASCI codes**

# Processor Specifics

- Processors
  - AMD Sledgehammer (Opteron)
  - 2.0 GHz
  - 64 Bit extension to IA32 instruction set
  - 64 KB L1 instruction and data caches on chip
  - 1 MB L2 shared (Data and Instruction) cache on chip
  - Integrated dual DDR memory controllers @ 333 MHz
  - Integrated 3 Hyper Transport Interfaces @ 3.2 GB/s each direction
- Node memory system
  - Page miss latency to local processor memory is <140 ns
  - Peak bandwidth of ~5.3 GB/s for each processor

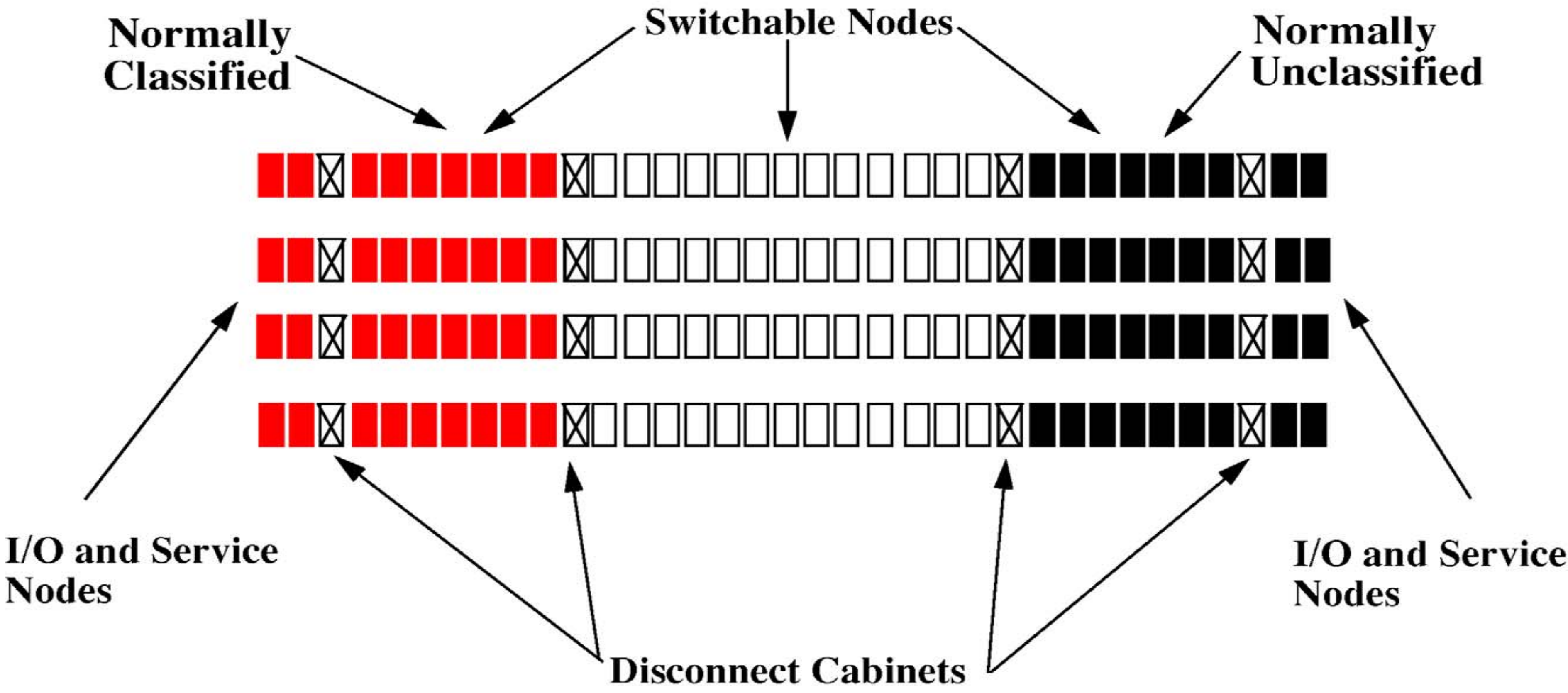
# Network and Network Topology

- Sandia has had very good experiences with the mesh topology
  - ASCI applications tend to be physical in nature. Mapping a 3D problem to a 3D machine preserves locality and maximizes us of fast “nearest neighbor” links.
  - Space-shared batch processing creates a communications locality that matches meshes very well
  - Works well with Red/Black switching
- Meshes look very promising for the future
  - The longest wire in the network determines performance
  - Meshes need no long wires

# Red Storm Topology

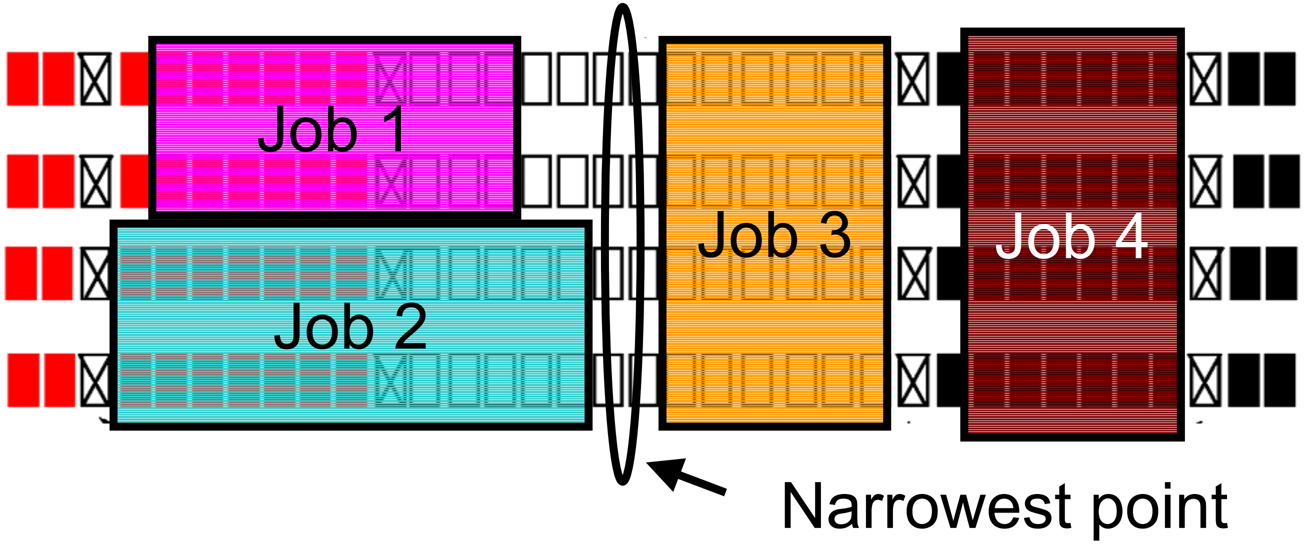
- Red Storm RFQ specifies a 3D mesh, Sandia and Cray concurred on specific topology
- Compute node topology:
  - 27 x 16 x 24 (x, y, z)
  - Mesh in x & y, torus in z
  - Red/Black split: 2,688 – 4,992 – 2,688
- Service and I/O node topology
  - 2 x 8 x 24 (x, y, z) on each end
  - 192 full bandwidth links to Compute Node Mesh (384 available)

# Red Storm Topology



# Advantages and Disadvantages

+ **Works well for space-shared batch processing**



- **An application crossing the narrowest point of the mesh has a "bisection bandwidth" constraint**
  - **Not sure Sandia has any of these**

# Interconnect Performance

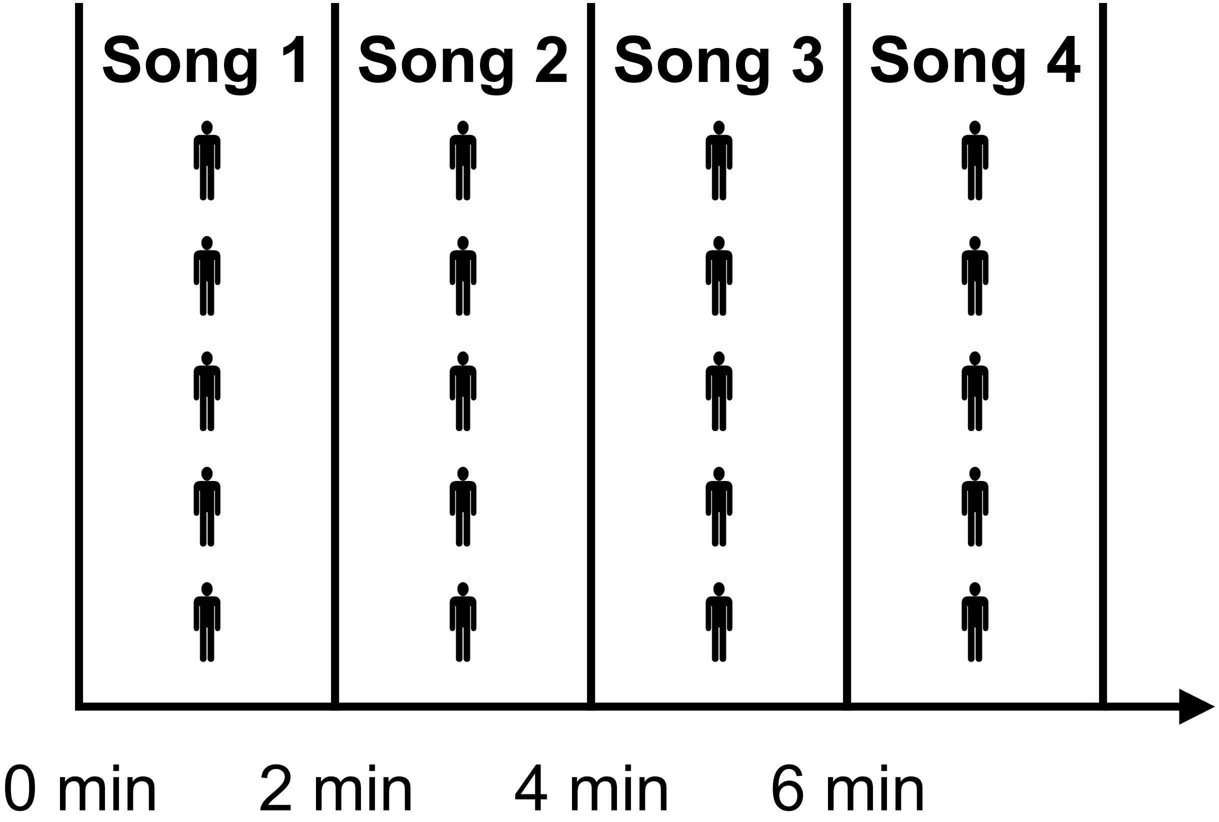
- Interconnect performance
  - MPI Latency <2 μs (neighbor), <5 μs (full machine)
  - Peak link bandwidth ~3.0 GB/s each direction (sustained 1.8 GB/s each direction)
  - Minimum bi-section bandwidth 1.5 TB/s
- I/O system performance
  - Sustained file system bandwidth of 50 GB/s for each color
  - Sustained external network bandwidth of 25 GB/s for each color

# Light Weight Kernel

- Sandia has had very good experiences with LWK
  - Sandia-University of New Mexico Operating System (SUNMOS)
  - Cougar
  - Puma
  - Now Catamount (tell story about name)
- Why?
  - Timing stability
  - Maturity

# LWK & Musical Rehearsal

- **N musicians Rehearsing 2 Minute Pieces**

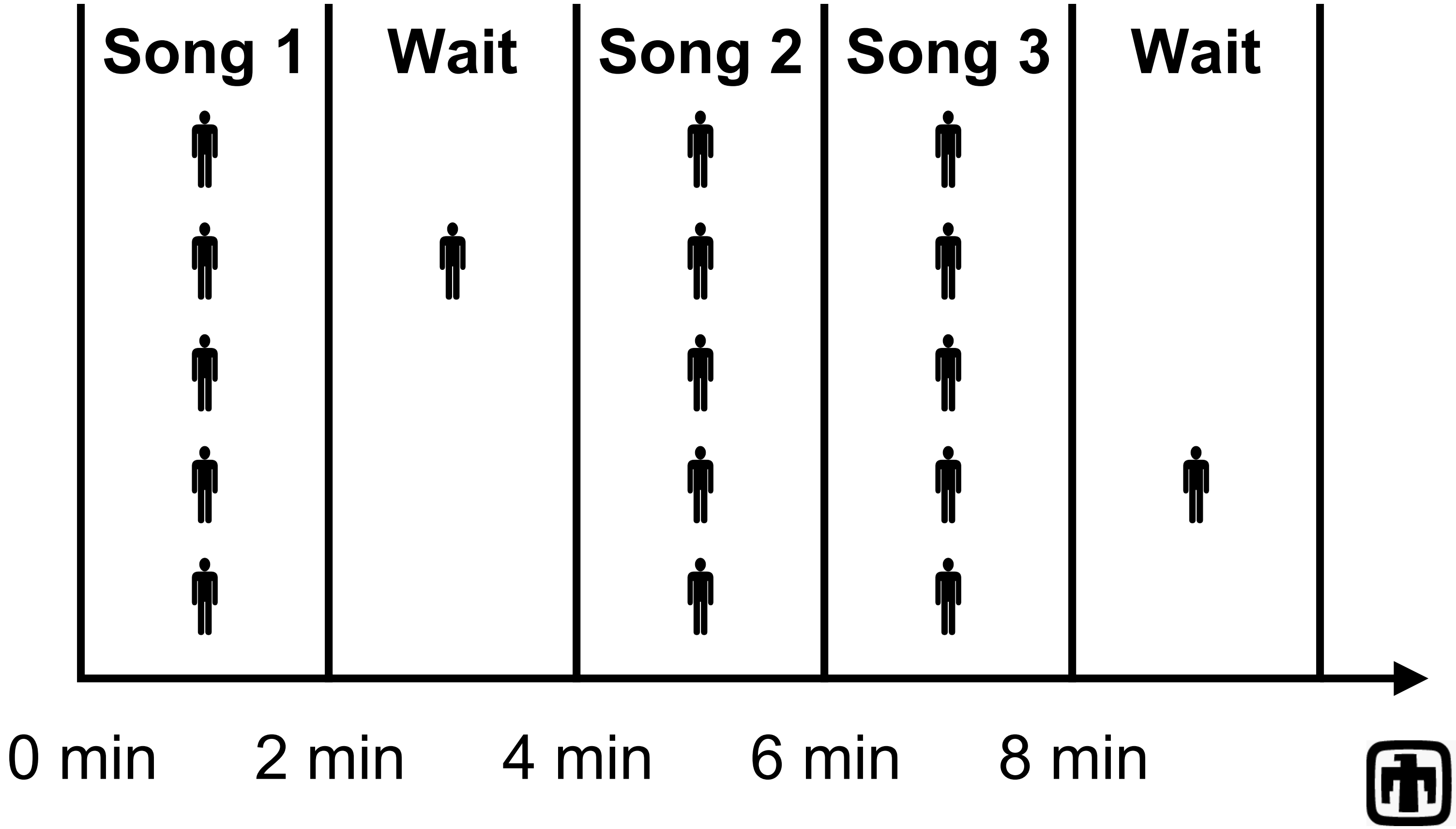


Song 1 Song 2 Song 3 Song 4

0 min 2 min 4 min 6 min

# Musical Rehearsal with Breaks

- 2 Minute Pieces with Asynchronous Breaks



# Breaks in MPP Systems Software

- Unix, Linux, any OS
  - Kernel memory allocation
  - TCP/IP backoff calculations
  - Routing tables
  - Clock synchronization
  - Scheduler
  - Etc., full list unknown, but has been extremely problematic with DOE labs

- Light Weight Kernel
  - [Intentionally blank]

# Run Time Impact of Unix Systems Services

- Say breaks take 50 μS and occur once per second
  - On one CPU, wasted time is 50 μs every second
    - Negligible .005% impact
  - On 100 CPUs, wasted time is 5 ms every second
    - Negligible .5% impact
  - On 10,000 CPUs, wasted time is 500 ms
    - Significant 50% impact

- Red Storm will be 10,000 CPUs, <u>but will not have asynchronous services</u>

# Red Storm Systems Software

- Operating Systems
  - LINUX on service and I/O nodes
  - LWK (Catamount) on compute nodes
  - LINUX on RAS nodes
- Run-Time System
  - Logarithmic loader
  - Node allocator
  - Batch system – PBS
  - Libraries – MPI, I/O, Math
- Parallel File System
  - Several file systems are being evaluated

# Reliability, Availability, and Serviceability

- Red Storm RFQ specifies 100 hour MTBI
  - You would take a PC back to Best Buy if it crashed every 4 days
  - However, Red Storm must be able to continue operating while nodes fail and get replaced just to meet this standard
- Red Storm will have a separate RAS network and system of 2500 Unix processors to manage the main machine
  - Will be able to pause running programs, reconfigure hardware, and continue

# RAS Network

- RAS Workstations
  - Separate and redundant RAS workstations for Red and Black ends of machine
  - System administration and monitoring interface
  - Error logging and monitoring for major system components including processors, memory, NIC/Router, power supplies, fans, disk controllers, and disks
- RAS Network: Dedicated Ethernet network for connecting RAS nodes to RAS workstations
- RAS Nodes
  - One for each compute board
  - One for each cabinet

# Red Storm Performance

**Peak of ~ 40 TF**

**Expected MP-Linpack performance >20 TF**

**Aggregate system memory bandwidth - ~55 TB/s**

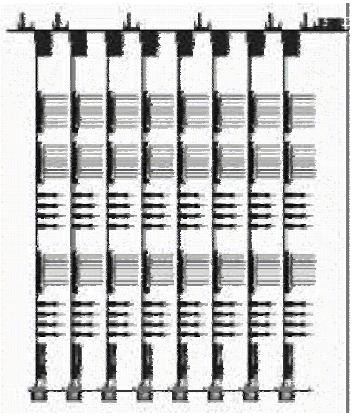
**Interconnect**

- **Aggregate sustained interconnect bandwidth > 100 TB/s**
- **MPI Latency - 2 µs neighbor, 5 µs across machine**
- **Bi-Section bandwidth ~2.3 TB/s**
- **Link bandwidth ~3.0 GB/s in each direction**

**Disk and External Network I/O**

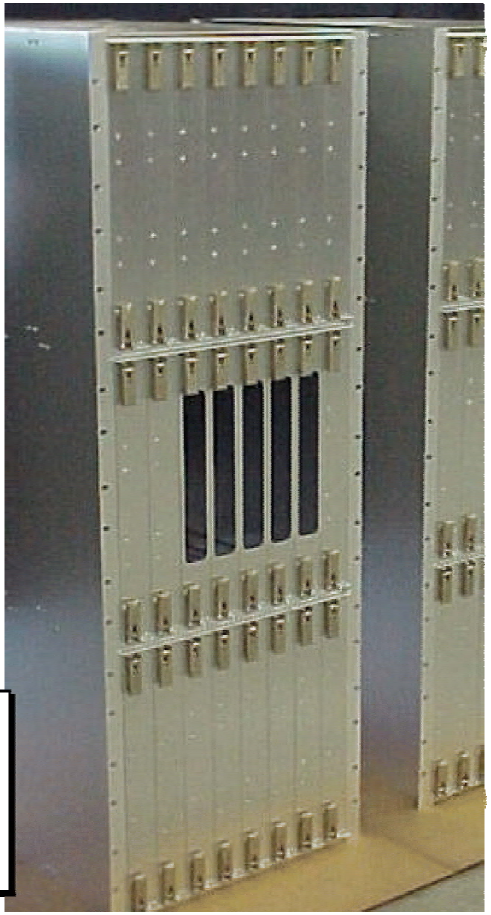
- **Sustained 50 GB/s each color parallel disk I/O**
- **Sustained 25 GB/s each color external network I/O**

# Red Storm Hardware Status



Card Layout - Top View

- 24 Boards
- 96 Operton™ Processors
- EMI containment
- Vertical Air Cooling



# Red Storm Hardware Status

Pre-Prototype Racks

# Comparison of ASCI Red and Red Storm

| | ASCI Red | Red Storm |
|---|---|---|
| Full System Operational Time Frame | June 1997 (Processor and Memory Upgrade in 1999) | August 2004 |
| Theoretical Peak (TF) | 3.15 | 41.47 |
| MP-Linpack Performance (TF) | 2.379 | >20 (est) |
| Architecture | Distributed Memory MIMD | Distributed Memory MIMD |
| Number of Compute Node Processors | 9,460 | 10,368 |
| Processor | Intel P II @ 333 MHz | AMD Opteron @ 2.0 GHz |
| Total Memory | 1.2 TB | 10.4 TB (up to 80 TB) |
| System Memory B/W | 2.5 TB/s | 55 TB/s |
| Disk Storage | 12.5 TB | 240 TB |
| Parallel File System B/W | 1.0 GB/s each color | 50.0 GB/s each color |
| External Network B/W | 0.2 GB/s each color | 25 GB/s each color |
| Interconnect Topology | 3-D Mesh (x, y, z)<br>38 X 32 X 2 | 3-D Mesh (x, y, z)<br>27 X 16 X 24 |

# Comparison of ASCI Red and Red Storm

| | ASCI Red | Red Storm |
|---|---|---|
| **Interconnect Performance** | | |
| MPI Latency | 15 µs 1 hop, 20 µs max | 2.0 µs 1 hop, 5 µs max |
| Bi-Directional Link B/W | 800 MB/s | 6.0 GB/s |
| Minimum Bi-section B/W | 51.2 GB/s | 2.3 TB/s |
| **Full System RAS** | | |
| RAS Network | 10 Mbit Ethernet | 100 Mbit Ethernet |
| RAS Processors | 1 for each 32 CPUs | 1 for each 4 CPUs |
| **Operating System** | | |
| Compute Nodes | Cougar | Catamount (Cougar) |
| Service and I/O Nodes | TOS (OSF1) | LINUX |
| RAS Nodes | VX-Works | LINUX |
| **Red Black Switching** | 2260 - 4940 - 2260 | 2688 - 4992 - 2688 |
| **System Foot Print** | ~2500 sq ft | ~ 3000 sq ft |
| **Power Requirement** | 850 KW | 1.7 MW |